DOCUMENT RESUME

ED 152 806 TM 006 967

AUTHOR Madaus, George F.
TITLE Background of the Project to Develop Guidelines and Standards for Educational Evaluation.
PUB DATE [Apr 77]
NOTE 7p.; Paper presented at the Annual Meeting of the American Educational Research Association (61st, New York, New York, April 4-8, 1977)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Curriculum Development; Educational Programs; Educational Research; *Educational Testing; Elementary Secondary Education; Evaluation Methods; Federal Programs; *Guidelines; Post Secondary Education; Professional Associations; *Program Evaluation; *Public Policy; *Standards

ABSTRACT

In the final stages of drafting the 1974 revision of the Standards for Educational and Psychological Tests, it became clear to the committee members that the various drafts of the revisions emphasized the assessment of individuals, rather than program evaluation. Because of its concern, the committee decided to finish the present draft and to recommend the preparation of a companion volume of standards for program evaluation to the three sponsoring bodies--the American Psychological Association, the American Educational Research Association, and the National Council on Measurement in Education. This brief report outlines the contents of the memorandum proposing the companion volume, and describes the initial actions taken by the three sponsoring associations on the basis of solicited evaluations. (Author/MV)

Background of the Project to Develop Guidelines and Standards for Educational Evaluation

George F. Madaus
Boston College

In the final stages of drafting the 1974 revision of the Standards for Educational and Psychological Tests it became clear to the committee members that the various drafts of the Standards revisions gave primary emphasis to standards that dealt with the assessment of individuals. While acknowledging problems in the use of tests in evaluation programs, curricula, therapeutic treatment, etc., the draft in its overall content and tone clearly focused on decisions affecting an examinee by an employer, personnel manager, guidance counselor, college admission officer, vocational counselor, clinician, teacher, etc.

A perusal of the letterheads over which individuals submitted reactions to the Final Draft of the revisions revealed that the bulk of the testimony came from individuals representing private firms or associations with interest in employment and personnel selection; from governmental agencies engaged in testing for employment or monitoring of employment practices; from associations representing minority groups; and from legal assistance groups with interest in unfair hiring practices.

Individuals associated with firms directly engaged in large scale program evaluation and/or policy research either submitted no testimony at all, or in the case of two firms, submitted comments which dealt with matters of individual assessment of criterion referenced measurement. That the focus and flavor of the revision, like that of the 1966 version, was primarily on the use of tests for the assessment of individuals was of course not surprising. The growing number of court cases involving the use of tests in civil service, private industry, and the civil rights area helped to shape the tone and thrust of the draft.

Thus the committee became concerned that while the draft document did have things to say to those using tests in program evaluation--the sense the document conveyed through its choice of language and examples was primarily one of a set of standards focused on the more traditional use of tests in individual assessment. Because of this concern the committee deliberated on whether to begin a new draft which would include additional standards for test use in the program evaluation domain, or whether to finish the present draft and to recommend to the three sponsoring bodies the preparation of a companion volume of standards for program evaluations. To help in these deliberations I was asked as a member of the committee to prepare a memorandum on the issues that might be addressed in a companion volume. At that time there was a pressing need to issue a revision of the 1966 Standards that dealt with issues dealing with the use of tests in selection and employment. After further deliberation the joint committee decided to recommend to APA, AERA and NCME that a companion volume be developed. The memorandum became the basis of a proposal for a companion volume and was circulated by the Board of Scientific Affairs of APA to approximately twenty individuals intimately concerned with large scale evaluation and policy research for their reactions as to its merits.

In the time remaining I will briefly outline the contents of the memorandum reactions to the proposal for a companion volume and the initial steps taken by APA, AERA and NCME on the basis of the solicited evaluations.

In outlining the contents of the memorandum I must of necessity be very brief and consequently I will for the most part omit specific examples used in the memorandum to illustrate the need for a companion volume. However, those of you wishing a copy of the memorandum can write to me at Boston College.

The 1974 Memorandum

After giving a brief history of the impetus behind the 1966 Standards and the subsequent growth after 1966 of program evaluation, the memo asked whether the proposed revision adequately dealt with issues related to test use in the following situations:

(a) The evaluation of the effectiveness of government sponsored educational interventions such as Head Start, Titles I, III, and VII of the Elementary And Secondary Education ACT (ESEA), Follow Through, METCO, etc.

(b) Educational research affecting public policy. This includes research affecting public policy. This includes research similar to that reported by Coleman (1966), Jencks et al, (1966), Jensen (1969), Herrnstein (1972), Armor (1972), etc.

(c) The formative and summative evaluation of large scale curriculum development projects (e.g., BSCS; Harvard Project Physics; the Aesthetic Education Program) and other educational products and packages produced by the Regional Laboratories. Programs like Sesame Street, ZOOM, The Electric Company, etc., developed under governmental and foundation grants could also be included under this category.

In addition to these three categories there were other developments related to the testing movement which were not addressed in the draft revision and which called for a consideration of possible standards of test use. These include the National Assessment Project, statewide needs assessments, performance contracting, criterion referenced testing, and accountability.

(A) Program Evaluation

Under the category of program evaluation the following issues were raised in the memorandum:

(1) The implications of using tests primarily designed to maximize individual differences for the evaluation of group performance.

(2) The implications of using norms based on individual performance for evaluating group performance.

(3) The proliferation of discourse on the properties of new tests (e.g., criterion referenced tests) cried out for definitive treatment. Standards relevant to the development, use and interpretation of alternatives to normed referenced tests needed to be developed.

(4) "New" tests apart, current test interpretation in program evaluation could benefit from new standards. Evaluations of federally funded programs lean heavily on the measurement of educational progress or growth. The increasing use of standardized achievement tests in the measurement of growth posed special problems not covered in present standardized test manuals or in the draft revision of the Standards. The implications of using various derived test metrics such as the GE to operationalize "growth" needed a detailed explication. Questions associated with problems of analyses of gain scores so crucial in longitudinal studies like Head Start andFollow Through were not considered in the draft revision.

It was argued that a new set of standards could bring together the best thinking on these important issues with a force and authority that would persuade companies performing evaluations, and test publishers in their manuals, to address themselves more carefully to problems of growth and gain.

(5) Another issue frequently encountered in program evaluation, not covered in the proposed draft was that of regression artifacts.

(B) Use of Tests in Public Policy Research

It seemed clear that public policy debate is often influenced by interferences drawn from test data. There is little doubt that studies by Jensen (1969), Herrnstein (1971), Armor (1972), Jencks et al, (1972), Coleman et al (1966) and the reanalysis of the Coleman data by Mosteller and Moynihan (1972), have all influenced the dialogue concerning educational policy, legislation and funding priorities for educational research. The validity of the conclusions from all of these studies rests primarily on inferences made from test data. A set of standards specifically geared to the use of tests in policy related research could have provided a framework within which a more rational debate of the merits of these studies could have taken place; just as the 1966 Standards acted as a framework within which the issue of test validity in discrimination cases was argued.

Coleman and Karweit (1972) offered three headings under which test results have been used to measure school and program effectiveness in policy research. The memorandum adopted these headings.

1. Depicting the level of functioning of Standards in an already existing program, school or school district.

Witness the yearly publication in such papers as the New York Times or the Boston Globe of school average reading scores at different grade levels. The revision of the Standards did not adequately deal with this type of test reporting or the misleading inferences to which these seemingly straightforward data lend themselves.

2. Describing the impact of a special program with a definite starting point.

The Wolff and Stein (1966) study of the impact of summer Head Start programs is a case in point under this heading.

3. Using test scores as "dependent variables" in research aimed at separating effects of student background from those of school environment.

Coleman's final category is applicable to his work. The IEA study (Husen, 1967) is another example in this category. There are many issues subsumed under this category. For example, in the Plowden Report, the question arises of how best to treat Time 1 achievement scores in an analysis which seeks to account for Time 2 achievement differences. Depending on the method used, the policy implementations differ considerably as has been outlined by Acland (1973):

Another issue involves which derived achievement score should be used in analysis and how mean GE scores should be computed. For example, an evaluator can compute the mean of a group of raw scores (or "standard scores") and convert that mean to another score metric such as a GE (or "standard score", whatever the case may be) for the group. Using the same set of scores, another evaluator could convert each raw score to a GE and compute the resulting mean of the GE distribution. The two mean GE's will not necessarily be the same.*

Since test validity refers to the accuracy of inferences from test scores, the inferences made from a group's actual test performance (i.e., number right) should not vary from one score to another.

Finally, there have been serious proposals to use test results to allocate funds for programs for the educationally disadvanraged. The implications of using various derived test scores in an allocation formula and as the basis for continuation of funds have not been fully thought through. New standards could inform the policy maker of this important and sensitive area.

(C) Evaluation of Large Scale Curriculum Development Projects

The intent of the evaluation process in R&D is to assure quality control for eventual products. However, the evaluation effort varies widely across development projects. The proliferation of literature on what constitutes good evaluation has not as yet been synthesized into a set of recognized standards. It was felt that a set of standards suitably promulgated would force educational entrepreneurs to reconsider their responsibilities for quality control.

The memo asked whether or not standards are needed (a) regarding the assessment procedures used to evaluate the various components of the product while under development (formative evaluation, (b), regarding the assessment procedures used to evaluate the products' effectiveness upon completion and release for sale (summative evaluation), (3) regarding the type and quality of the information that the developer/publisher must provide the consumer. The standards called for in the last instance are analogous to those governing test manuals; and (d) regarding the procedures used to determine the educational as well as the cost effectiveness of the particular package under evaluation vis a vis other packages purporting to accomplish the same objectives. This type of comparative curriculum evaluation encounters many of the difficulties discussed in section (A) above.

*The difference can range from one to four tenths of a G.E., based on simulated data using the California Test Battery.

If evaluation is really an integral part of curriculum development, if its function is to inform that development and protect the consumer, then a consideration of standards relative to the evaluation process in R&D, AERA, and NCME appear to be in order.

D. Other Issues

It was felt that several other developments in testing during the past several years needed scrutiny in terms of standards. For these, no attempt was made to develop the related measurement issues. Instead, the memo suggested that there are important public policy issues relevant to test usage in the areas of statewide assessment, accountability, performance contracting and the National Assessment Project. A new committee would need to scrutinize these areas with an eye toward suggesting standards where needed.

Review of the Memorandum

In brief then, these were the issues described in the memorandum. The evaluation of the memorandum was favorable. Most of the reviewers felt that there was a need for standards or at least guidelines in the areas outlined. Further, it was generally felt that the time was ripe to begin this undertaking. Many individuals cited additional issues beyond test use that the new committee should also address. These included:

--ethical/philosophical issues
--control of bias
--evaluation of evaluation
--essentials to be included in an evaluation report
--centrality of valuing process
--evaluator's relationship to programs evaluated, funders of evaluation, audiences, and general public.
--analysis techniques
--standard related to tactics of consciously or unconsciously reducing the quality of pretests (test-administrator malingering)
--standard relating to preserving data for reanalysis (freedom of information) and protecting participant privacy
--experimental design considerations
--presentation of results
--ground rules for research
--responsibilities in serving different audiences
--experimental design and statistical control
--matrix sampling testing procedures
--legal questions
--minimum specifications for evaluation contracts
--conceptualization of evaluation
--state and federal regulations
--organization and management concerns

Further, the reviewer cited several excellent examples of work already under way which would be relevant to the development of a new set of standards for program evaluation. These included work by Tyler, Scriven, Messick, Stake, Lumsdaine, Novick, Freeman, Bederman, Wholy, Eash, Stufflebeam, Coleman, Palmer, Sanders, and ETS.

On the basis of these favorable reviews the three agencies which sponsored the 1974 Revision of the Standards appointed a committee composed of Egon Guba of AERA, Don Campbell, Robert Senn and Henry Reichen of APA, Ron Carver, Dan Stufflebeam and myself from NCME to decide upon the next steps to be taken in the development of a companion volume. That Committee met in Chicago in May of 1975, and I think this point is best taken up by our next speaker, Dan Stufflebeam.

References

Acland, H.D. Social determinants of educational achievement: an evaluation and criticism of research. Ph.D. thesis submitted to Oxford University, 1973.

Armor, J. The evidence on busing. The Public Interest, No. 28, (Summer 1972), pp. 90-126.

Coleman, J.S., et al. Equality of educational opportunity. Washington, D.C.: U.S. Government Printing Office, 1966.

Coleman, J.S. and Karweit, N. Information systems and performance measures in schools. Englewood Cliffs, N.J.: Educational Technology Publishers, 1972.

Herrnstein, R. IQ. The Atlantic Monthly, Sept. 1971, 43-64.

Husen, T. International study of achievement in mathematics. New York: John Wiley and Sons, 1967.

Jencks, C. et al. Inequality: a reassessment of the effect of family and schooling in America. New York: Basic Books, 1972.

Jensen, A. R. How much can we boost IQ and scholastic achievement? Harvard Educational Review, 39, 1969, 1-23.

Wolff, M. and Stein, A. A comparison of children who had Head Start, Summer 1965 with their classmates in kindergarten. New York: Ferkauf Graduate School of Education, Yeshiva University, 1966 (mimeo).

Mosteller, F. and Moynihan, D.P. (Eds.) On equality of educational opportunity. New York: Random House, 1972.